UNIVERSIDAD AUTONOMA DE MADRID

Escuela Politécnica Superior

19|20

Trabajo fin de grado

Modelado y predicción de generación de residuos sólidos usando enfoque de aprendizaje automático

Óscar Gómez Borzdynski

Escuela Politécnica Superior
Universidad Autónoma de Madrid
C\Francisco Tomás y Valiente nº 11

excelencia Campus Internacional UAM CSIC

www.uam.es

UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR

Doble Grado en Ingeniería Informática y Matemáticas

TRABAJO FIN DE GRADO

Modelado y predicción de generación de residuos sólidos usando enfoque de aprendizaje automático

Autor: Óscar Gómez Borzdynski
Tutor: Ángel Sánchez Calle
Ponente: David Renato Domínguez Carreta

junio 2020

Óscar Gómez Borzdynski
Modelado y predicción de generación de residuos sólidos usando enfoque de aprendizaje automático

Óscar Gómez Borzdynski
C\ Francisco Tomás y Valiente Nº 11

IMPRESO EN ESPAÑA – PRINTED IN SPAIN

Gracias a mi familia por el apoyo incondicional. Gracias a mi pareja por la paciencia y las veces que se ha leído este trabajo. Gracias a mi tutor por el conocimiento que me ha brindado y sus grandes ideas.

Cuanto más sabes, más te das cuenta de que no sabes nada

Sócrates

Resumen

Con el creciente interés por la degradación del medio ambiente debido a la cantidad de residuos generados por la humanidad, decidimos realizar este trabajo para buscar un método de predicción de generación de residuos sólidos que permita la redistribución de recursos tanto económicos como políticos y proteger el medio ambiente. Utilizamos datos de la OCDE, publicados en su página web de estadísticas, con este propósito. Probamos varios algoritmos englobados dentro del Aprendizaje Automático, así como dos formas de generar el conjunto de variables, obteniendo resultados bastante buenos. Finalmente, relacionamos el error respecto al modelo y los valores reales con las características de algunos países.

Palabras clave

Generación de residuos sólidos, Aprendizaje automático, Algoritmos de predicción, Series Temporales de Datos, Modelado de Información, Maquinas de Vectores Soporte, Potenciación del gradiente, Proceso gaussiano, Redes neuronales, Informes de la OCDE

Abstract

With the growing interest for the environmental degradation among politicians and general public due to the increasing amount of solid waste generated, we decide to make this investigation looking for a method to predict the waste to be generated by some countries in order to allow them to redistribute some resources to protect the environment. We used the OECD data published in their statistical page to achieve this. We tried different Machine Learning algorithms, as well as two ways of generating the dataset obtaining really satisfactory results. Finally, we compare the error between the prediction and real data with some features of some countries.

Keywords

Solid waste generation, Machine Learning, Prediction algorithms, Time series of data, Information modeling, Support Vector Machines (SVM), Gradient boosting, Gaussian process, Neural networks, OECD reports

Índice

1 Introducción **1**

1.1 Estado del arte 1

1.2 Objetivo 2

1.3 Método 2

1.4 Estructura 3

2 Algoritmos **5**

2.1 Support Vector Regressor 5

2.2 Gradient Boosting Regressor 6

2.3 Gaussian Process Regressor 6

3 Preparación de los datos **9**

3.1 Variables geográficas 9

3.2 Variables económicas 10

3.3 Variables demográficas 11

4 Experimentos **15**

4.1 Experimento 1: SVR sin conocimiento del pasado 15

4.2 Experimento 2: GBR sin conocimiento del pasado 16

4.3 Experimento 3: GPR sin conocimiento del pasado 19

4.4 Experimento 4: SVR con conocimiento del pasado 21

4.5 Experimento 5: GBR con conocimiento del pasado 21

4.6 Experimento 6: GPR con conocimiento del pasado 23

4.7 Experimento 7: LSTM y DNN 26

4.8 Análisis de los datos mediante SVM sin conocimiento del pasado 27

5 Conclusiones **31**

Bibliografía **34**

1

Introducción

La generación de basura y su deposición es un tema muy investigado actualmente, pudiendo encontrar publicaciones que estudian la generación de basuras en países subdesarrollados [1], comparación de métodos en la predicción de basura municipal [2] o la predicción de basura generada semanalmente [3]. En la primera de las publicaciones se comprueba que en países subdesarrollados las metrópolis generan más basura por habitante que las zonas rurales, así como en las zonas costeras los residuos tenían una composición diferente a los presentes en las zonas de interior. En la segunda se comprueba una gran hetereogeneidad entre métodos utilizados con el propósito de predecir las basuras municipales, lo que motiva una comparación de algoritmos bajo las mismas condiciones en los datos. En la última de ellas se prueban distintos métodos basados en Redes Neuronales para buscar mejores resultados reduciendo la cantidad de variables de entrada.

Además, con el cambio climático en boca de todos, esta rama de investigación está más en auge ahora que nunca, existiendo estudios en los que se busca el impacto medioambiental producido por la generación de basura y sus distintos métodos de procesamiento [4]. El *Machine Learning* se justifica como técnica en este campo debido a su desarrollo a pasos agigantados también en otros campos, usándose para la predicción meteorológica [5], conducción autónoma [6] o incluso en traducción [7].

1.1. Estado del arte

Se han realizado trabajos previos en los que se trata de predecir los desechos generados mediante algoritmos de Aprendizaje Automático. En 2013, *Abbasi et al.* [8] exploraron la utilización de una híbrido entre vectores de soporte y transformación Wavelet para predecir la generación de basura municipal en Teherán. Para el SVM, probaron con distintos *kernels* así como con distintas funciones de error y demostraron que los algoritmos de *Machine Learning* pueden ser utilizados para realizar estas predicciones.

Antanasijević et al. [9] probaron a realizar el mismo proceso utilizando una red neuronal artificial para predecir la generación de basuras en ciertos países europeos. Utilizaron una red neuronal con únicamente 3 capas y 3 variables de entrada. En este caso se llegó a la conclusión de que algunos

de estos países, como Serbia, carecía de datos suficientemente precisos para obtener resultados satisfactorios.

En 2017, *Masebinu et al.* [10] manifestaron la importancia de los datos de basuras municipales aportados para el entrenamiento de una red neuronal artificial, comprobando que el tamaño de la población influye en esta predicción y detectando la diferencia de calidad entre los datos aportados por países africanos y europeos.

Posteriormente se utilizaron algoritmos similares para la predicción de residuos en regiones concretas, como en *Purcell et al.* [11], donde se centraron en la región de Dublín. En dicha publicación se comprueba la dependecia de factores socio-económicos con la generación de residuos, apreciando la diferencia entre distintos barrios con diversos niveles de vida e ingresos.

1.2. Objetivo

En este trabajo vamos a intentar predecir la basura generada por los países de la Organización para la Cooperación y el Desarrollo Económicos (OCDE), organización que pretende conseguir prosperidad y bienestar para las personas de los países miembros mediante nuevas políticas. Usaremos los datos recopilados por la OCDE en su sección estadística para predecir los desechos generados.

Con ello buscamos un posible método para obtener la cantidad de residuos sólidos generados por un país, lo que, en caso de aplicarse, permitiría a dicho país redistribuir recursos con intención de mejorar la calidad de vida de los habitantes de dicho país. Así mismo, los métodos estudiados se aplican para obtener predicciones globales relativas al conjunto de países de la OCDE.

1.3. Método

Utilizaremos varios algoritmos de inteligencia artificial, como regresión mediante máquinas de vectores de soporte (SVR) [12], regresión por *boosting* del gradiente (BGR) [13] o procesos Gaussianos (GP) [14] para predecir la basura generada en un año por estos países. Por último, comprobaremos si el *Deep Learning* [15] tiene relevancia para este problema mediante la utilización de red neuronal secuencial (con un cierto número de capas ocultas) y otra con introducción de capas *Long-Short Term Memory* (LSTM) [16]. Además probaremos con distintas formas de representar los datos para tratar de mejorar el rendimiento de los algoritmos. Emplearemos las librerías Scikit-Learn [17] y Keras [18] con *backend* en Tensorflow [19].

Posteriormente evaluaremos la variación de rendimiento de uno de los algoritmos en función de la cantidad y tipo de datos socio-económicos usados para el entrenamiento de los modelos y sobre qué país se aplica, tratando de ver qué países generan una cantidad diferente de basuras a la que predice

el algoritmo. De esta manera podemos realizar una separación de países en función de la relación entre estos factores y la producción de residuos.

Como datos de entrada del algoritmo utilizaremos estadísticas socio-económicas como el número de habitantes, el nivel de estudios o la edad de los habitantes, entre otras variables. Estos datos han sido obtenidos de diversas fuentes, pero mayoritariamente se obtuvieron de la página de datos estadísticos de la OCDE. Analizaremos estos datos brevemente viendo qué características socio-económicas influyen más en la producción de residuos.

Este trabajo viene motivado por la falta de predictores aplicados sobre datos de conjuntos grandes de países. La OCDE engloba países de varios continentes y las diferencias culturales presentes pueden influenciar en gran medida estas predicciónes. En caso de obtener buenos resultados podrían aplicarse técnicas de planificación urbana y de manejo de residuos para minimizar el impacto en el medio ambiente de estas basuras. Se han realizado estudios relacionando algunos factores socio-económicos con la composición de las basuras [20–22]. La unión de ambas técnicas podría además aportar mayor detalle en la predicción, permitiendo una redistribución de recursos más eficiente.

1.4. Estructura

Este trabajo se organiza en 5 capítulos. Inicialmente, en el capítulo 2 se explicarán brevemente los algoritmos que se usarán a lo largo del trabajo. En el capítulo 3 se hablará de los datos socio-económicos seleccionados y sus orígenes, así de cómo se han transformado para su uso en los algoritmos. En el capítulo 4 se tratarán todos los experimentos realizados, sus resultados y propuestas de mejora de cada uno. Para finalizar, en el capítulo 5 se presentarán las conclusiones obtenidas de todo el trabajo, así como un trabajo futuro y posibles modificaciones.

2

ALGORITMOS

A lo largo de esta sección usaremos la siguiente notación: $\{(x_i, y_i)\}_{i=1}^{n} \subset \mathcal{X} \times \mathcal{Y}$, donde x_i representa las variables de entrada al algoritmo, en nuestro caso $x_i \in \mathbb{R}^m$. $y_i \in \mathbb{R}$ representa el dato conocido que se pretende aproximar, n es el número de casos conocidos y m es el número de variables de entrada.

2.1. Support Vector Regressor

Comenzamos definiendo la regresión mediante máquinas de vectores de soporte (SVR) [23] [12] lineales.

Este algoritmo se basa en encontrar una función:

$$f(x) = w^\top x + b, W \in \mathbb{R}^m, b \in \mathbb{R}$$

tal que cada elemento y_i de $\mathcal{Y}$ se sitúe a una distancia máxima ϵ de $f(x_i)$. Además se busca minimizar $\frac{1}{2}\|w\|^2$. Por ello podemos escribir el problema como:

$$\begin{aligned} &\text{minimizar } \frac{1}{2}\|w\|^2 \\ &\text{sujeto a: } \left|y_i - w^\top x - b\right| \leq \epsilon \end{aligned}$$

Ahora bien, esto sólo lo podemos hacer si dicha función existe con el ϵ decidido. En caso de que esto no sea posible debemos definir los errores γ_i, por lo que reescribimos:

$$\begin{aligned} &\text{minimizar } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\gamma_i \\ &\text{sujeto a: } \left|y_i - w^\top x - b\right| \leq \epsilon + \gamma_i \\ &\gamma_i \geq 0 \end{aligned}$$

En caso de utilizar un *kernel* (al no ser los datos linealmente separables en el espacio original), se realiza la siguiente modificación:

$$\text{minimizar } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\gamma_i$$
$$\text{sujeto a: } |y_i - K(x,w) - b| \leq \epsilon + \gamma_i$$
$$\gamma_i \geq 0$$

donde K es la función del *kernel*.

Debido a ello en este algoritmo tendremos 2 parámetros, ϵ y C. El primero indica el error que estamos dispuestos a asumir sin penalizar al algoritmo, el segundo indica qué importancia se le da a los errores respecto a los pesos de la función.

2.2. Gradient Boosting Regressor

Este algoritmo [24] [13] utiliza una combinación de regresores más débiles en forma de árboles de decisión para aproximar la función $f(x_i) = y_i$. El primer parámetro de este algoritmo es el número de estimadores (árboles) que entrenar, M. Por tanto, tratamos de obtener la siguiente igualdad:

$$f_{m+1}(x_i) = f_m(x_i) + h_m(x_i) = y_i \quad \forall x_i \in \mathcal{X}$$

Es decir, buscamos $h_m(x_i) = y_i - f_m(x_i)$, el estimador $m+1$ será un árbol de decisión que aproximará la diferencia entre el estimador anterior y el resultado deseado. El segundo parámetro que considerar en este algoritmo es la profundidad de este árbol. Esto implica el número de separaciones en el espacio que realizará el estimador. En caso de tomar una profundidad p, tendremos 2^p divisiones.

Para comenzar el algoritmo, debemos definir $f_0(x_i) = \bar{\mathcal{X}}$. El último parámetro es el ratio de aprendizaje, r, que regula la importancia que se le da a cada estimador. Por ello el resultado final de nuestro algoritmo es:

$$f(x_i) = \bar{\mathcal{X}} + r\sum_{m=1}^{M} f_m(x_i)$$

2.3. Gaussian Process Regressor

Un proceso Gaussiano [14] es un método Bayesiano que devuelve una distribución de predicciones para cada entrada. Dada esa distribución podemos obtener la medición en el punto de mayor probabilidad junto con la confianza de dicha predicción mediante el análisis de la varianza.

Es un modelo no paramétrico basado en distribuciones Gaussianas multivariables, definidas por un vector de medias μ y su matriz de covarianza Σ. El vector μ representa el valor esperado de la distribución. La matriz Σ siempre es simétrica y semi-definida positiva, la diagonal de ésta contiene las varianzas de cada variable aleatoria mientras que el resto de elementos indica la covarianza entre variables.

Utilizamos una distribución normal dado que, gracias al Teorema del Límite Central, la agregación de distintas distribuciones se aproxima a una Gaussiana. Cabe destacar que la dimensión de la distribución Gaussiana será igual al número de puntos de entrenamiento, por lo que para problemas con muchas muestras es muy costoso computacionalmente.

Usualmente utilizamos *kernels* para calcular Σ permitiendo tratarla como una matriz de distancias entre variables aleatorias. Esto también permite introducir no-linealidad al modelo.

El entrenamiento de este modelo se realiza mediante la introducción de puntos conocidos en el conjunto de variables aleatorias y condicionando nuestra distribución a dichos puntos. Supongamos que al introducir nuestros puntos Y a nuestras variables aleatorias X obtenemos una distribución normal de forma:

$$P_{X,Y} = \begin{bmatrix} X \\ Y \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_{X,X} & \Sigma_{X,Y} \\ \Sigma_{Y,X} & \Sigma_{Y,Y} \end{bmatrix}\right)$$

Podemos condicionar dichas variables utilizando la regla de Bayes multivariable:

$$P(X|Y) = \mathcal{N}\left(\mu_X + \Sigma_{X,X}\Sigma_{Y,Y}^{-1}(Y - \mu_Y), \Sigma_{X,X} - \Sigma_{X,Y}\Sigma_{Y,Y}^{-1}\Sigma_{Y,X}\right)$$

Esto se cumple para una función perfecta, pero no es el caso habitual cuando se pone en práctica cualquier modelo. Para ello introducimos un ruido Gaussiano $\epsilon = \mathcal{N}(0, \Psi^2)$

Utilizaremos la siguiente probabilidad conjunta:

$$P_{X,Y} = \begin{bmatrix} X \\ Y \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_{X,X} & \Sigma_{X,Y} \\ \Sigma_{Y,X} & \Sigma_{Y,Y} + \Psi^2 I \end{bmatrix}\right)$$

Terminada esta introducción, el parámetro que modificaremos será la elección del *kernel*, permitiendo adaptarnos a cómo se presentan los datos en la distribución.

3

Preparación de los datos

En este capítulo describiremos los datos utilizados a lo largo de este trabajo. En primer lugar, describiremos los datos de basuras de la OCDE. Todos estos datos han sido obtenidos de la página *OECD Statistics* (https://stats.oecd.org/), donde se encuentran divididos en distintos *datasets* categorizados por área. Corresponden a 28 años (1990-2017) para los 43 países que pertenecen a la organización. Añaden además una serie de divisiones de la propia OCDE, como el total o sólo la parte europea. Los datos se obtienen en formato CSV y serán transformados a Parquet para su procesamiento en Pandas. Las basuras están divididas en 6 grupos:

- **Basura municipal**, se refiere a la basura generada por los municipios de ese país; engloba comercios, viviendas, instituciones públicas (hospitales, escuelas...) y desechos orgánicos procedentes de zonas verdes como jardines privados o parques públicos. No incluye desechos de construcción.
- **Basura doméstica**, se refiere sólo a la basura generada por las viviendas de los municipios.
- **Basura recuperada**, se refiere a la basura que ha sido transformada en algún elemento con un potencial beneficio económico o ecológico. Engloba la generación de biocombustible, reciclado, compostaje o reutilizado.
- **Basura reciclada**, se refiere al procesado de la basura para obtener un material utilizable. El reciclado en el mismo producto también está englobado en este apartado.
- **Basura compostada**, descomposición aeróbica o anaeróbica de residuos con intenciones de usarse posteriormente.
- **Basura desechada**, basura que no es aprovechada para ningún otro fin; engloba dos procedimientos, el primero es mediante algún tratamiento, como la incineración o la transformación química sin resultado utilizable. El segundo es el desecho, depositado de basura en vertederos, almacenamiento permanente o vertido al mar.

Para predecir los valores previamente mencionados, hemos elegido una serie de variables socioeconómicas de dichos países. En el caso de no conocer el dato para un año concreto, hemos realizado alguna técnica de interpolación que se especificará para cada caso.

Hemos dividido estas variables en 3 grupos: geográficas, demográficas y económicas.

3.1. Variables geográficas

En este grupo se recogen las variables que dependen del país como zona geográfica. Son variables que no tienen gran variación a lo largo del tiempo.

La primera variable es el **área del país**. Esta variable suele verse asociada a recursos naturales, espacio disponible para la construcción, espacio disponible para desechar residuos, etc. Solo se ve afectada por cambios en las fronteras de los países o modificación en los procedimientos para su cálculo. Estos datos presentaban 3 mediciones en el periodo utilizado. Dado que esta variable no presenta mucha variación a lo largo del tiempo decidimos utilizar el último valor conocido para las posiciones no conocidas.

Esta variable presenta grandes diferencias entre países. El valor mínimo se da para Luxemburgo con $2,590$ y el máximo para la suma de la OCDE con un valor de $36,570,179$. La diferencia entre ellas es de orden 10^4 .

La segunda variable es la **proporción de área construida**. Se ve afectada por factores humanos, pero no suele presentar grandes variaciones en un periodo de un año. Se ve asociada a la cantidad de población, industria y desarrollo de un país. Esta variable, al igual que la anterior, presenta solo 3 mediciones. Dado que no es una variable con gran variación se sigue el mismo procedimiento para rellenar los datos no conocidos.

En este caso, debido a que se trata de un porcentaje, el valor está acotado entre 0 y 1, por lo que no vamos a encontrar diferencias de gran magnitud. Aun así, vemos un máximo de $16{,}96\,\%$ en los Países Bajos y un mínimo de $2\,\%$ para Islandia.

3.2. Variables económicas

En este grupo de variables se recogen los factores económicos que definen un país. Suelen verse relacionados con el desarrollo de un país, así como su estabilidad y capacidad de crecimiento. Todos los datos fueron obtenidos en la página de estadísticas de la OCDE.

Como primera variable de este grupo encontramos la **mediana de ingresos**. Esto define el nivel de ingresos de una persona media, así como su nivel de vida. Los datos se encontraban en moneda local, por lo que lo hemos convertido a dólares para poder equipararlos. En este caso podemos ver mucha variación. El mínimo se encuentra en la India en 2004, con $245{,}98\$$; el máximo lo encontramos en Luxemburgo en 2013, con $43863{,}07\$$. Podemos advertir que la población de Luxemburgo gana prácticamente 180 veces más que los habitantes de la India.

La segunda variable económica será la **Renta per Cápita**, que define el desarrollo económico de un país y que, desde 1970 [25], se ha desacoplado de la mediana. Por ello pese a parecer medidas muy similares indican factores muy diferentes, siendo éste un factor de desarrollo industrial o tecnológico, mientras que la mediana indica el nivel de vida de los ciudadanos. En este caso también convertimos a dólares desde la moneda local para poder compararlas. El máximo lo encontramos en Luxemburgo en 2017 con $126156{,}45\$$; el mínimo es de $1{,}40\$$ en Colombia en 1990.

En tercer lugar, incluiremos el **porcentaje del producto bruto procedente del turismo**, esta variable se introduce debido a que los turistas suelen reciclar menos que los habitantes habituales, además de un consumo superior de bienes. En este caso también podemos fijarnos en los valores extremos, siendo el mínimo $1{,}18\,\%$ en Polonia y un máximo de $14{,}62\,\%$ en Portugal.

3.3. Variables demográficas

En este último bloque incluimos variables que refieren a la población del país. Estos estadísticos demográficos aportan información acerca de la composición de la población, tanto en edad como en estudios y la posible implicación que tienen con el cuidado del medio ambiente, el reciclaje y las medidas de conservación. Todos los datos fueron obtenidos en la página de estadísticas de la OCDE.

La primera variable que incluimos es la **Población total**. Un aumento en la población supone un aumento en la demanda de bienes, provocando una mayor generación de desechos. En este caso el país más poblado es China, con 1409517000 habitantes en 2017; el país menos poblado es Islandia, con 296734 habitantes en 2005.

En segundo lugar, incluimos el **Porcentaje de población sin estudios de secundaria**, indicativo del nivel de estudios de la población. Esta variable se ha añadido debido a una posible relación entre el nivel de estudios y la concienciación con el medio ambiente y, por consiguiente, la voluntad de reciclaje. En este caso el máximo se encuentra en Turquía con un $86{,}36\,\%$ de habitantes sin la educación secundaria completada; el mínimo se encuentra en Rusia con un $4,76\,\%$ de habitantes sin secundaria.

Para finalizar incluiremos los grupos de edad, es decir, el porcentaje de personas que se encuentran en un rango determinado de edad. Según [26], esta variable influye en la conciencia respecto al medio ambiente, por lo que utilizaremos los siguientes grupos:

- **Menores de 20 años**, según [26] los jóvenes son el grupo que menor conciencia de cuidado del medio ambiente presentan.
- **Mayores de 50 años**, edad con mayor conciencia.
- **Mayores de 65 años**, edad en la que, según [26] comienza a disminuir dicha conciencia.
- **Mayores de 85 años**, edad con menor conciencia.

Con todo esto, podemos ver que disponemos de 11 variables para predecir 6 tipos de basuras distintos.

Para posteriores experimentos aportaremos cierta información del pasado. Para ello, añadimos las basuras generadas el año anterior para que el algoritmo pueda basarse en los valores del año previo. De esa manera, esperamos que apoyándose en ese valor, el algoritmo pueda buscar sólo las variaciones.

Aparte de este dato, añadiremos los siguientes del año anterior: **porcentaje construido**, **porcen-**

taje de población sin estudios de secundaria, **población total** y **porcentaje del producto bruto procedente del turismo**. Para buscar las variaciones, aportaremos las diferencias de estas variables frente al año actual.

Como primer análisis de estos datos realizaremos una matriz de covarianzas, mostrada en la figura 3.1, donde apreciamos una correlación muy cercana entre las basuras de un año y las del año anterior. También existe una relación próxima entre la cantidad de basura que genera un país con su extensión y el porcentaje construido. Estas relaciones tienen cierto sentido lógico. Por otra parte, comprobamos que se cumplen las suposiciones de [26] ya que la relación del grupo de edad mayor de 85 con los deshechos es mayor que el resto de grupos. Estos datos se pueden apreciar con más detalle en la gráfica 3.2. Por último, destacamos la relación inversa entre la población sin estudios de secundaria y la generación de residuos, indicando que cuanto más educada es una población más residuos genera.

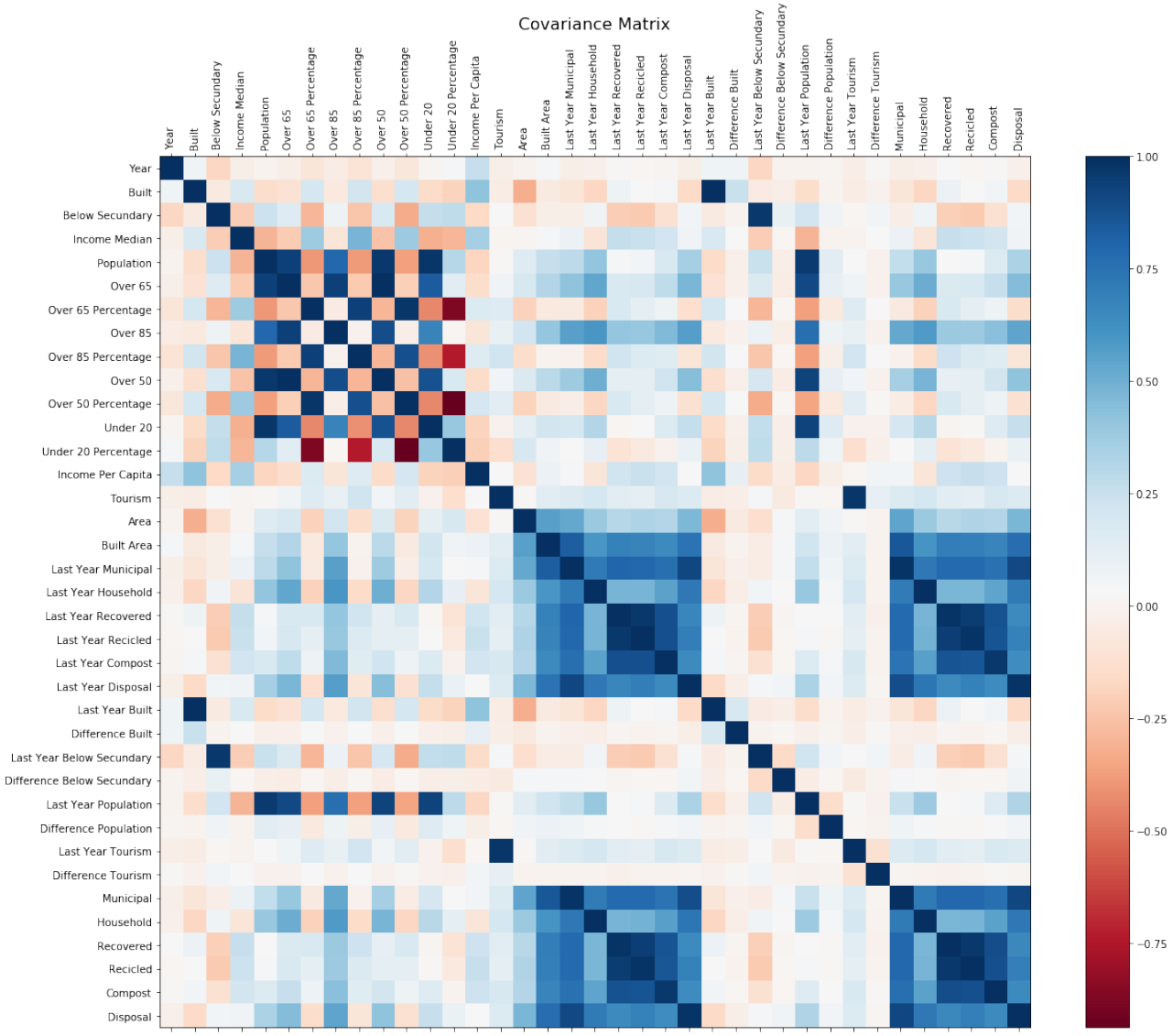

Figura 3.1: Matriz de covarianzas entre todas las variables: azul indica relación directa, rojo relación inversa. Los valores claros indican una relación leve, blanco indicaría independencia.

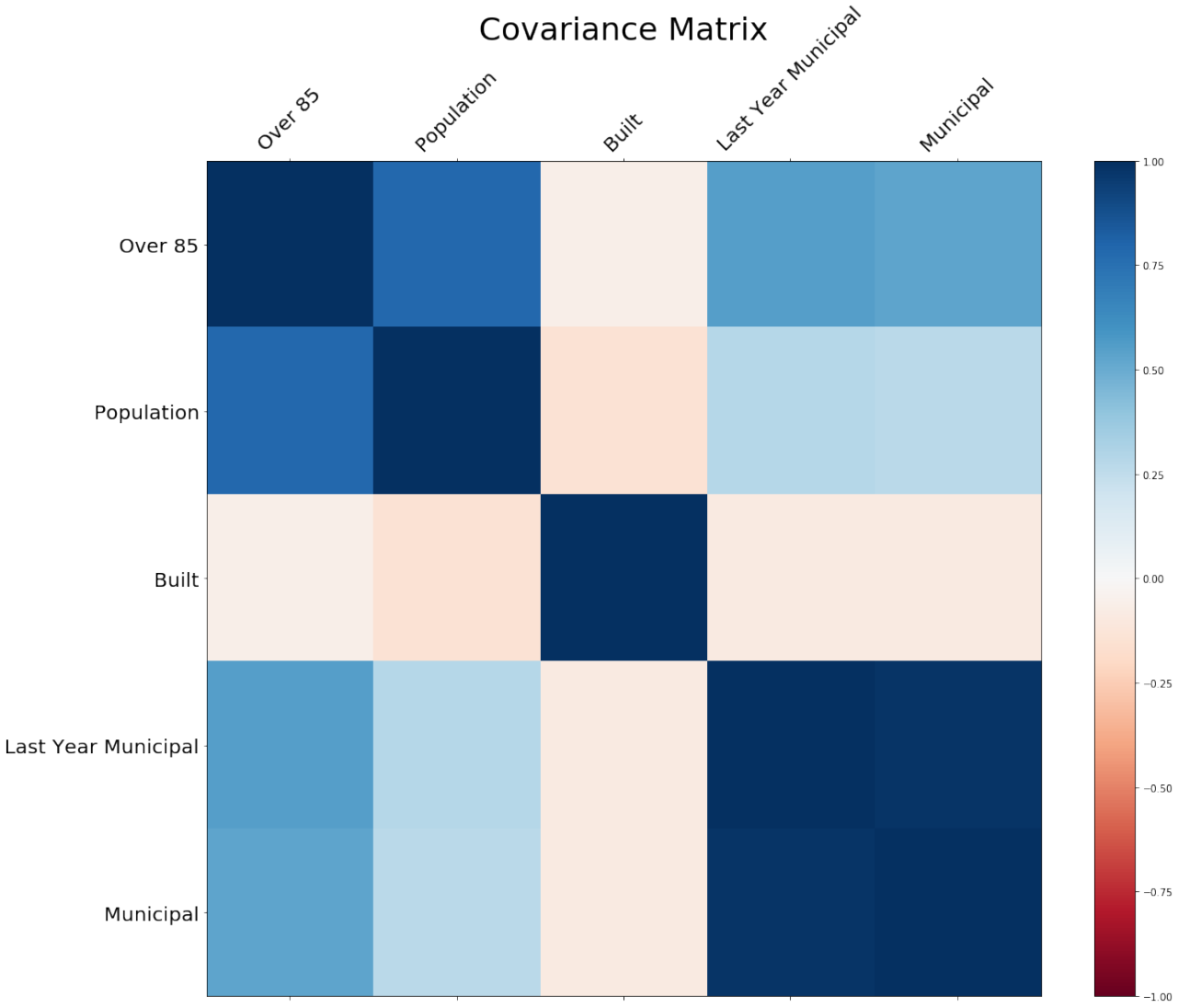

Figura 3.2: Matriz de covarianzas entre algunas variables seleccionadas: azul indica relación directa, rojo relación inversa. Los valores claros indican una relación leve, blanco indicaría independencia.

4

EXPERIMENTOS

En este capítulo vamos a tratar los experimentos realizados sobre los datos generados en el capítulo anterior. Usaremos Python3 sobre un Jupyter notebook. Las librerías principales son Pandas [27] y Scikit-Learn [17]. En todos los experimentos usaremos España como ejemplo para visualizar cómo predice el modelo entrenado.

4.1. Experimento 1: SVR sin conocimiento del pasado

Como primera aproximación, comenzaremos con una regresión mediante SVM [23] con los datos generados sin información del pasado.

En primer lugar, separaremos los datos en un conjunto de entrenamiento y otro de prueba. Hemos decidido utilizar un corte temporal en vez de una muestra aleatoria ya que se asemejaría más a un entorno real, donde conocidos los datos de este año y todo lo anterior se predice la generación de residuos del año actual. Escogemos el año 2014 como separación, dejando desde 1990 a 2013 (24 años) para entrenamiento y desde 2014 a 2017 (4 años) para prueba. De esta manera dejamos un $86\,\%$ de los datos para entrenamiento y un $14\,\%$ para prueba.

Dado que las variables presentan mucha variación entre ellas, vamos a normalizarlas mediante la técnica Min Max [28]. Esto es necesario porque, como se ha visto en el capítulo anterior, tenemos variables que toman valores de orden 10^9 y otras que como máximo toman valor 1.

Con todo ello procedemos a entrenar el modelo. Utilizamos un SVR para cada variable a predecir. Para elegir los parámetros del algoritmo realizamos una búsqueda entre todas las posibles combinaciones de los siguientes:

- **Kernel:** lineal ó RBF
- **C:** 0.5, 1, 5, 10, 50 ó 100
- ϵ**:** 0.1, 0.01 ó 0.001

Entre ellos, las combinaciones que mejores resultados obtienen se expresan en la tabla 4.1.

Como vemos, en todos los casos tenemos un coeficiente R^2 superior a 0.9, este coeficiente se

Variable a predecir	Kernel	C	ϵ	R^2
Municipal	RBF	50	0.001	0.9911
Doméstica	RBF	100	0.001	0.9611
Recuperada	RBF	10	0.001	0.9877
Reciclada	RBF	10	0.001	0.9780
Compostada	RBF	5	0.001	0.9760
Desechada	RBF	10	0.001	0.9915

Tabla 4.1: Parámetros de los SVR utilizados en el experimento 4.1 y sus coeficientes R^2.

acota superiormente por 1 e indica la bondad de la predicción respecto a los valores reales basándose en la varianza de los datos, si su valor es cercano a 1 significa que el modelo ajusta bien los datos. Visualizamos en la figura 4.1 cómo predice el modelo la generación de basuras en España. Para ello, dibujamos el valor real para toda la serie y el valor predicho para los últimos años. La separación entre los datos de entrenamiento y prueba se realiza con una línea vertical.

Como podemos apreciar, los datos aportados en España parecen comenzar en 2008 excepto para la basura municipal, donde tenemos datos desde 1995 y para la basura doméstica, que no presenta ningún dato más que el generado por defecto para el informe de la OCDE. En cuanto a nuestro modelo, vemos que según nos alejamos de los datos conocidos el error aumenta progresivamente. Esto no sucede sólo para España sino para todos los países, siendo la media del error absoluto progresivamente mayor en función de la distancia a la separación. Por otra parte, vemos que la basura municipal es la más abundante y al mismo tiempo es una de las que mejor resultado global obtiene.

4.2. Experimento 2: GBR sin conocimiento del pasado

En este segundo experimento utilizaremos un G*radient Boosting Regressor* como modelo. Como en el caso anterior, utilizaremos el primer *dataset* generado en el capítulo 3, por lo que no aportaremos información del pasado. La separación de los datos se realiza igual que en el experimento anterior y utilizaremos el mismo método de normalización debido al buen resultado obtenido anteriormente.

Para obtener los mejores hiperparámetros del modelo realizamos el mismo proceso que en el caso anterior, probando las combinaciones de las siguientes opciones:

- **Ratio de aprendizaje:** 0.001, 0.005, 0.01, 0.05, 0.1 ó 0.5
- **Número de árboles:** 50, 100, 200, 500 ó 1000
- **Profundidad máxima:** 2, 3, 5, 10, 20, 30, 50 ó100

Los mejores resultados se expresan en la tabla 4.2 y, utilizándolos para predecir en España, obtenemos las gráficas de la figura 4.2.

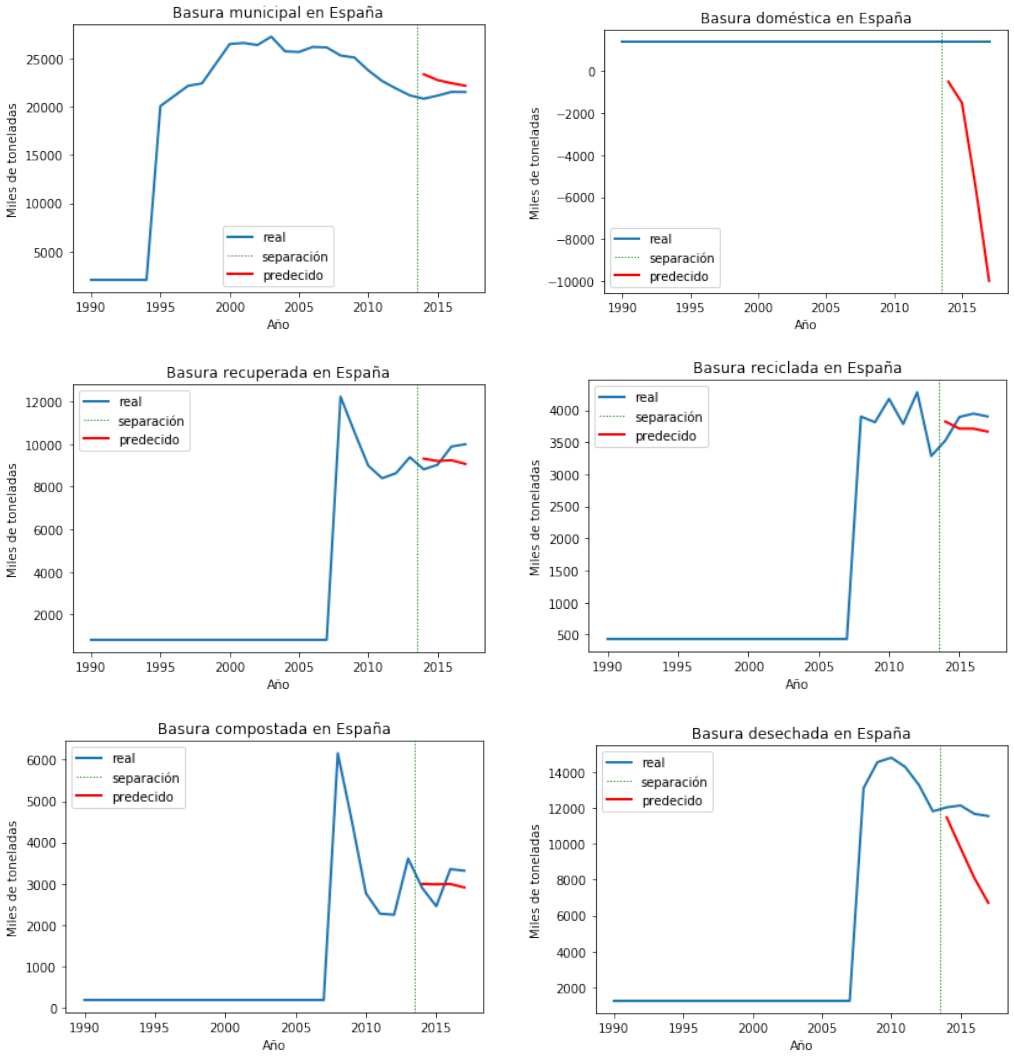

Figura 4.1: Valores de basura reales (azul) y predicho (rojo) para los 6 tipos definidos en España utilizando los SVR definidos en la tabla 4.1.

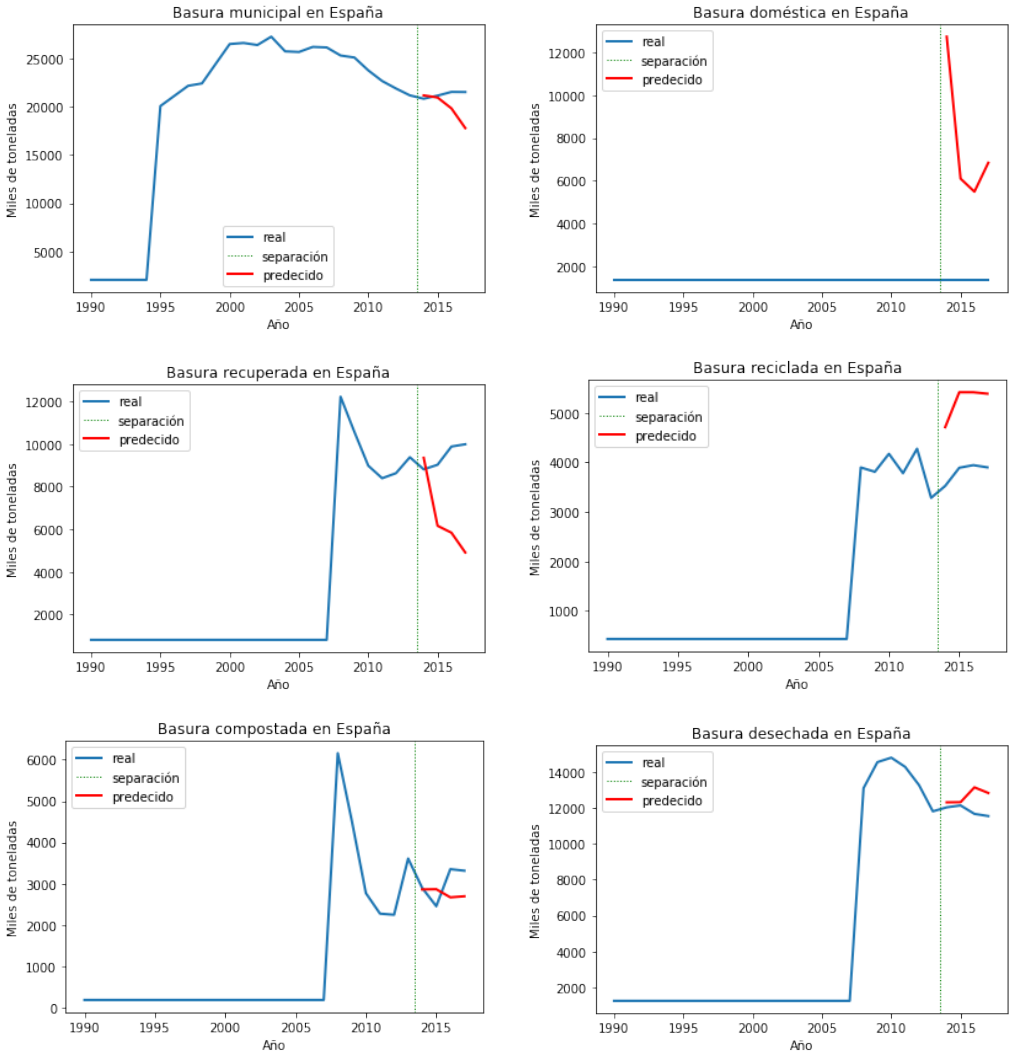

Figura 4.2: Valores de basura reales (azul) y predicho (rojo) para los 6 tipos definidos en España utilizando los GBR definidos en la sección 4.2.

Al igual que en el experimento anterior, nos fijamos también en los coeficientes R^2 de cada predictor en la tabla 4.2.

Variable a predecir	Profundidad máxima	Ratio de aprendizaje	estimadores	R^2
Municipal	3	0.05	1000	0.9383
Doméstica	2	0.1	500	0.8686
Recuperada	5	0.5	1000	0.9863
Reciclada	10	0.5	1000	0.9899
Compostada	5	0.1	1000	0.9830
Desechada	5	0.5	50	0.9817

Tabla 4.2: Parámetros de los GBR utilizados en el experimento 4.2 y sus coeficientes R^2.

Podemos apreciar que este algoritmo obtiene mejores resultados para la basura reciclada y la compostada, mientras que para el resto de tipos pierde algo de rendimiento. El caso más llamativo es para la basura doméstica, donde su coeficiente disminuye en gran medida. En las gráficas se puede ver que para España, las predicciones para las basuras recuperada y reciclada son bastante peores que en el experimento anterior.

4.3. Experimento 3: GPR sin conocimiento del pasado

En este caso utilizamos un proceso Gaussiano como regresor. Al igual que en los experimentos previos, utilizaremos datos sin conocimiento del pasado, sólo introduciendo datos del año a predecir. Separamos los datos en 2014 y utilizamos la técnica de normalización Min Max.

Volvemos a utilizar una búsqueda de hiperparámetros para elegir un *kernel* entre los siguientes:

- RationalQuadratic() + WhiteKernel()
- DotProduct() + WhiteKernel()
- DotProduct() + RationalQuadratic()
- RBF() + WhiteKernel()
- RBF() + DotProduct()
- RBF() + RationalQuadratic()

La mejor combinación para cada variable se expresa en la tabla 4.3; para España obtenemos los resultados de la gráfica 4.3.

Vemos que en este caso sucede lo mismo que en los dos experimentos anteriores. A medida que el dato a predecir se aleja del último año de datos con el que se entrena, el error aumenta. Esto se aprecia fácilmente en la predicción para basura reciclada en España.

Los coeficientes R^2 de este modelo se exponen en la tabla 4.3.

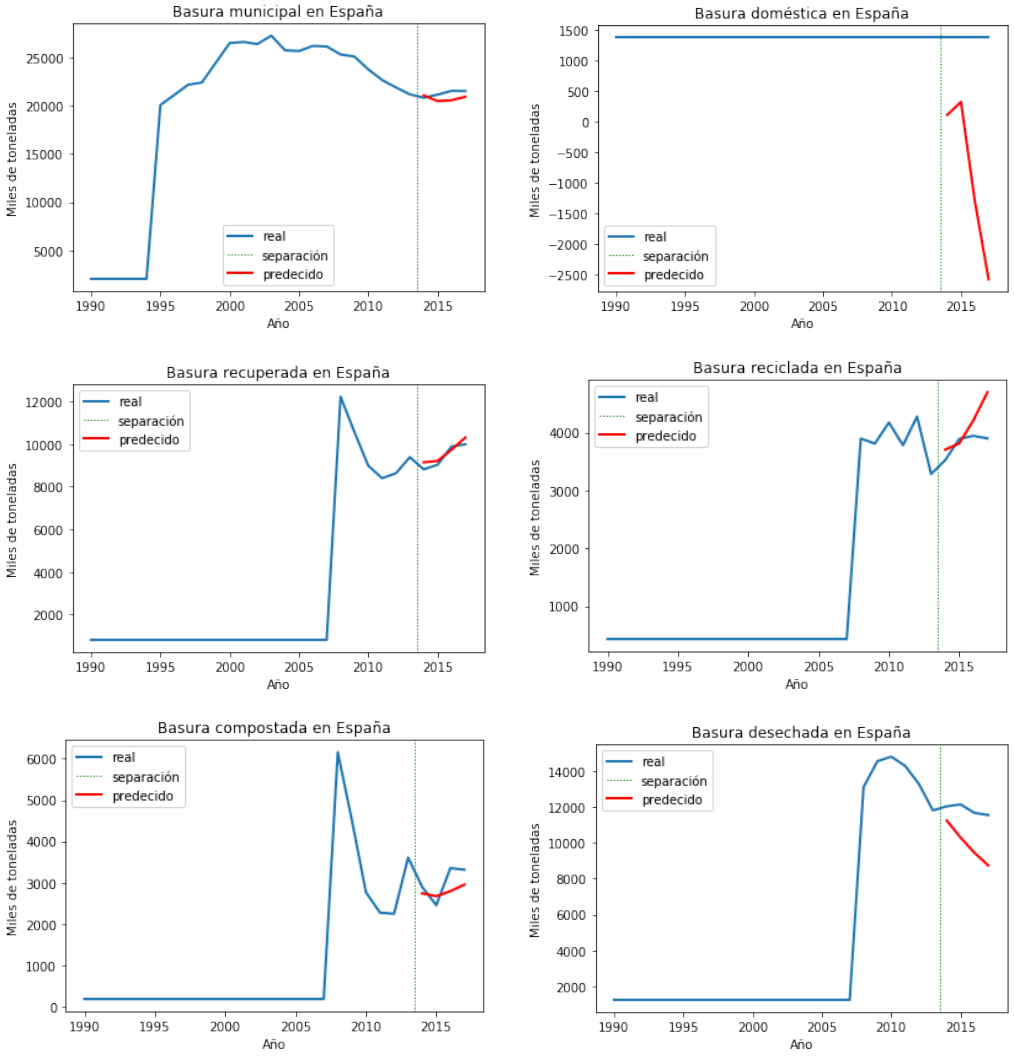

Figura 4.3: Valores de basura reales (azul) y predicho (rojo) para los 6 tipos definidos en España utilizando los GPR definidos en la sección 4.3.

Variable a predecir	Kernel	R^2
Municipal	RationalQuadratic + WhiteKernel	0.9786
Doméstica	RationalQuadratic + WhiteKernel	0.9693
Recuperada	RationalQuadratic + WhiteKernel	0.9732
Reciclada	RationalQuadratic + WhiteKernel	0.9499
Compostada	RationalQuadratic + WhiteKernel	0.9732
Desechada	RationalQuadratic + WhiteKernel	0.9793

Tabla 4.3: Parámetros de los GPR utilizados en el experimento 4.3 y sus coeficientes R^2.

4.4. Experimento 4: SVR con conocimiento del pasado

A partir de este experimento, empezamos a incluir información del pasado mediante el segundo conjunto de entrenamiento descrito en el capítulo 3, es decir, añadimos algunas variables al conjunto de datos utilizado en los experimentos anteriores. Entre estas variables encontramos, entre otras, las basuras del año anterior o la variación de tamaño de la población.

Escogemos el mismo año para separar el conjunto de entrenamiento y prueba, de esa manera, queda el periodo desde 1992 a 2013 ($85\,\%$) como conjunto de entrenamiento. El periodo de 2014 a 2017 ($15\,\%$) se utiliza para probar el rendimiento del modelo.

En este caso, volvemos a utilizar un SVR con la técnica de normalización Min Max, al igual que en el experimento 4.1. Los hiperparámetros se seleccionan de entre el mismo conjunto que en el experimento 4.1 y los resultados se expresan en la tabla 4.4.

Variable a predecir	Kernel	C	ϵ	R^2
Municipal	Lineal	500	0.001	0.9998
Doméstica	Lineal	100	0.001	0.9985
Recuperada	Lineal	1000	0.001	0.9991
Reciclada	Lineal	500	0.001	0.9989
Compostada	Lineal	1	0.001	0.9986
Desechada	Lineal	100	0.001	0.9997

Tabla 4.4: Parámetros de los SVR utilizados en el experimento 4.4 y sus coeficientes R^2.

Los resultados de este experimento son mucho mejores que los del experimento 4.1, pese a ser el mismo algoritmo. Por ello, podemos afirmar que aportar información del año anterior es muy beneficioso para este problema tal y como se esperaba basándonos en la matriz de covarianzas representada en el capítulo 3. Entonces, probamos de nuevo los algoritmos utilizados en los experimentos 4.2 y 4.3.

En la tabla 4.4 vemos que las gráficas responden al coeficiente R^2, teniendo un error menor y por lo tanto situándose más cerca la predicción del modelo de los valores reales. En la basura compostada vemos un comportamiento de copia del año anterior, es decir, el algoritmo parece devolver el valor para dicha basura proporcionado en la variable del año previo, sin embargo, comprobando otros países, hemos visto que sólo sucede en unos pocos y no es un fenómeno general.

4.5. Experimento 5: GBR con conocimiento del pasado

Tal y como hemos visto en el experimento 4.4, añadir la información del año anterior parece ayudar a los modelos, en este caso probaremos un GBR con dicha información. Para este caso, seleccionamos los parámetros entre el mismo conjunto que el experimento 4.2, mostrando los que mejores resultados

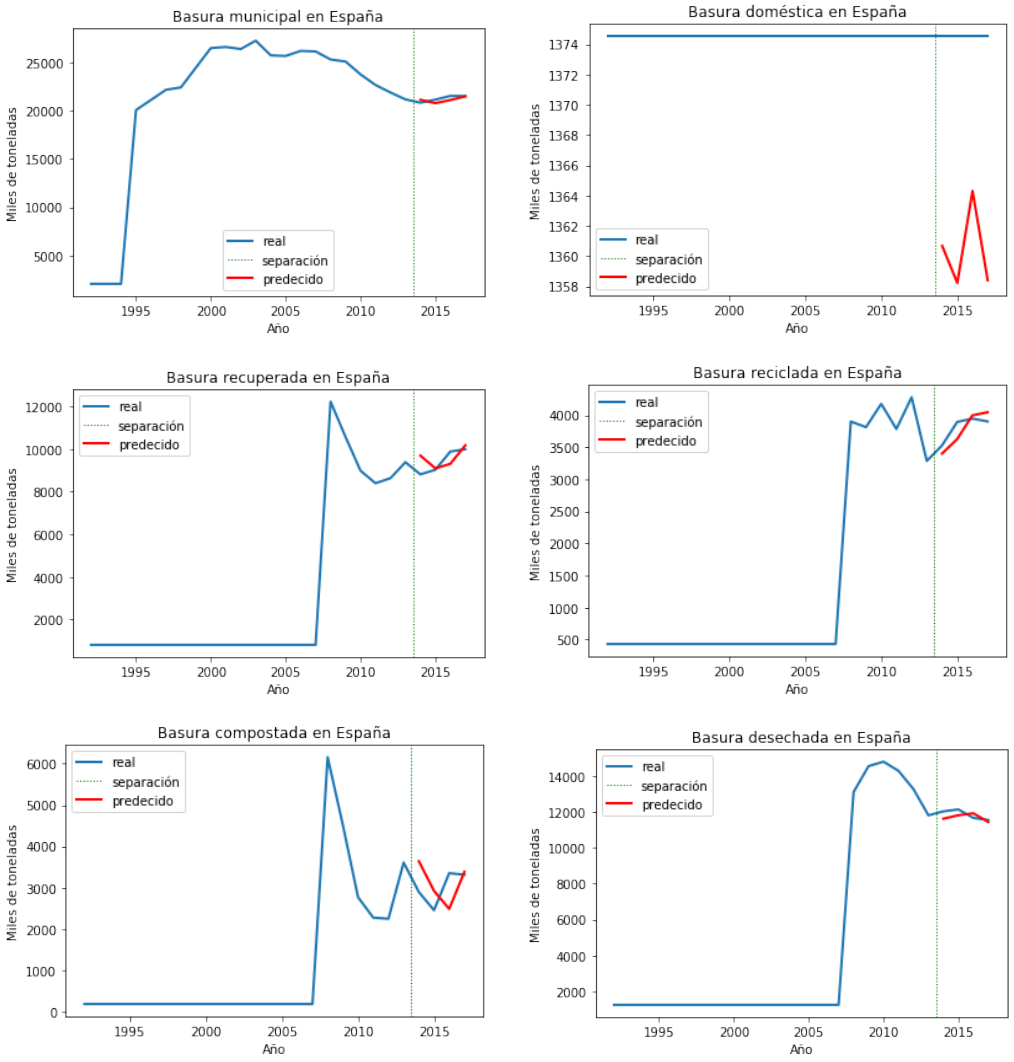

Figura 4.4: Valores de basura reales (azul) y predicho (rojo) para los 6 tipos definidos en España utilizando los SVR definidos en la tabla 4.4.

obtienen en la tabla 4.5.

Variable a predecir	Profundidad máxima	Ratio de aprendizaje	estimadores	R^2
Municipal	3	0.5	200	0.9942
Doméstica	10	0.1	1000	0.9980
Recuperada	30	0.5	500	0.9780
Reciclada	2	0.01	1000	0.9702
Compostada	20	0.5	1000	0.9696
Desechada	2	0.01	1000	0.9932

Tabla 4.5: Parámetros de los GBR utilizados en el experimento 4.2 y sus coeficientes R^2.

Podemos ver que en algunos casos el coeficiente R^2 es menor que en el experimento 4.2, sin embargo, en algunos casos sí obtenemos un rendimiento mejor. Estos casos son la basura municipal, la doméstica y la recuperada, aunque esta última en menor medida. En este caso nos podemos fijar en la gráfica para la basura reciclada, cuya aproximación es mucho mejor que para el experimento 4.2 pese a que su coeficiente es menor.

4.6. Experimento 6: GPR con conocimiento del pasado

En este apartado repetimos el experimento 4.3 con el mismo *dataset* usados en los experimentos 4.4 y 4.5. Los *kernels* seleccionados entre el conjunto utilizado en el experimento 4.3 se expresan en la tabla 4.6.

Variable a predecir	Kernel	R^2
Municipal	RationalQuadratic + WhiteKernel	0.9987
Doméstica	DotProduct + RationalQuadratic	0.9982
Recuperada	RationalQuadratic + WhiteKernel	0.9986
Reciclada	RationalQuadratic + WhiteKernel	0.9975
Compostada	RationalQuadratic + WhiteKernel	0.9930
Desechada	RationalQuadratic + WhiteKernel	0.9982

Tabla 4.6: Parámetros de los GPR utilizados en el experimento 4.6 y sus coeficientes R^2.

En este caso los coeficientes R^2 son muy cercanos a 1, lo que indica que nuestro modelo realiza predicciones muy cercanas a los valores reales. Esto se puede observar en las gráficas de la figura 4.6, donde los valores predichos son realmente cercanos a los reales y tienen formas similares a las del experimento 4.4.

Una vez hemos realizado todos estos experimentos pasamos a resumirlos en la tabla 4.7, donde hemos resaltado en negrita los clasificadores con mejores resultados.

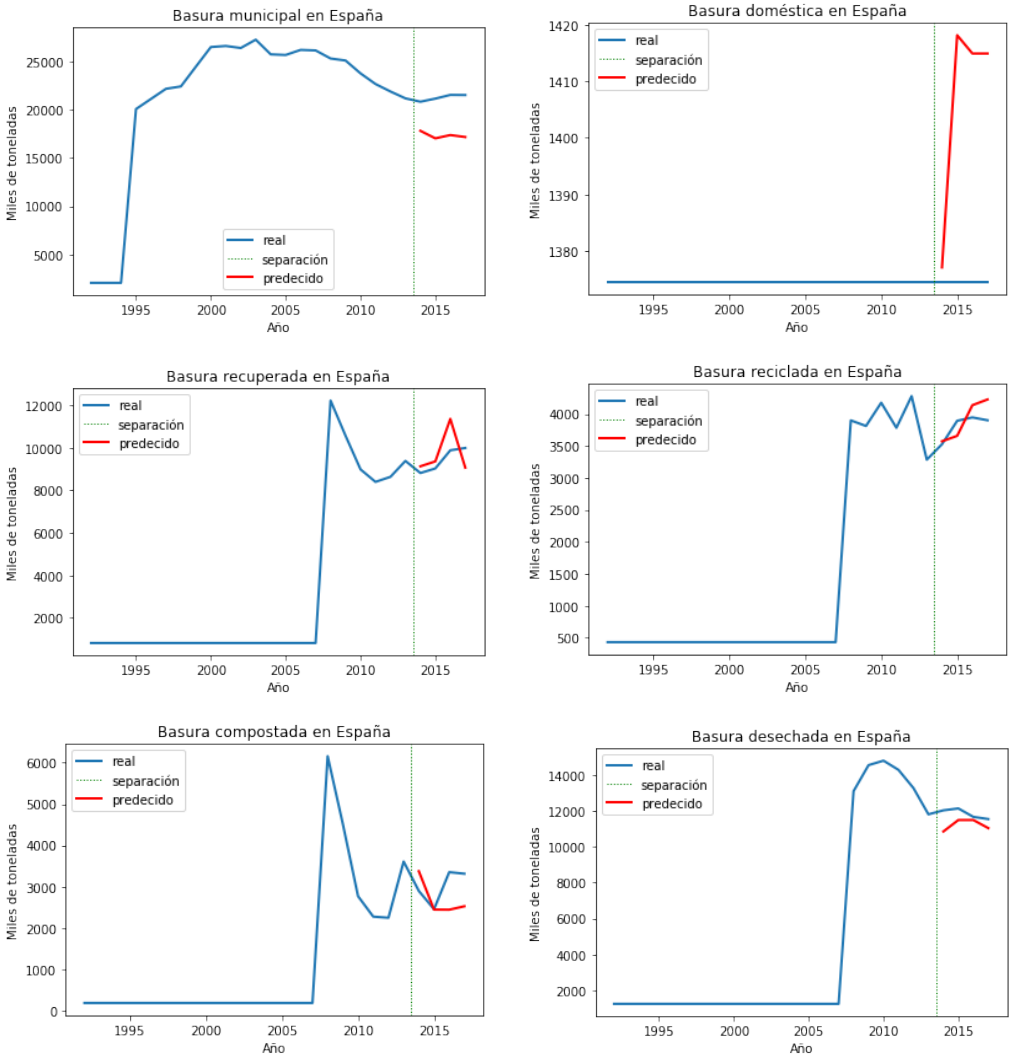

Figura 4.5: Valores de basura reales (azul) y predicho (rojo) para los 6 tipos definidos en España utilizando los GBR definidos en la tabla 4.5.

	4.1	4.2	4.3	4.4	4.5	4.6
Municipal	0.9911	0.9383	0.9786	**0.9998**	0.9942	0.9987
Doméstica	0.9611	0.8686	0.9693	**0.9985**	0.9980	0.9982
Recuperada	0.9877	0.9863	0.9732	**0.9991**	0.9780	0.9986
Reciclada	0.9780	0.9899	0.9499	**0.9989**	0.9702	0.9975
Compostada	0.9760	0.9830	0.9732	**0.9986**	0.9696	0.9930
Desechada	0.9915	0.9817	0.9793	**0.9997**	0.9932	0.9982

Tabla 4.7: Coeficientes R^2 de los experimentos realizados en el capítulo 4.

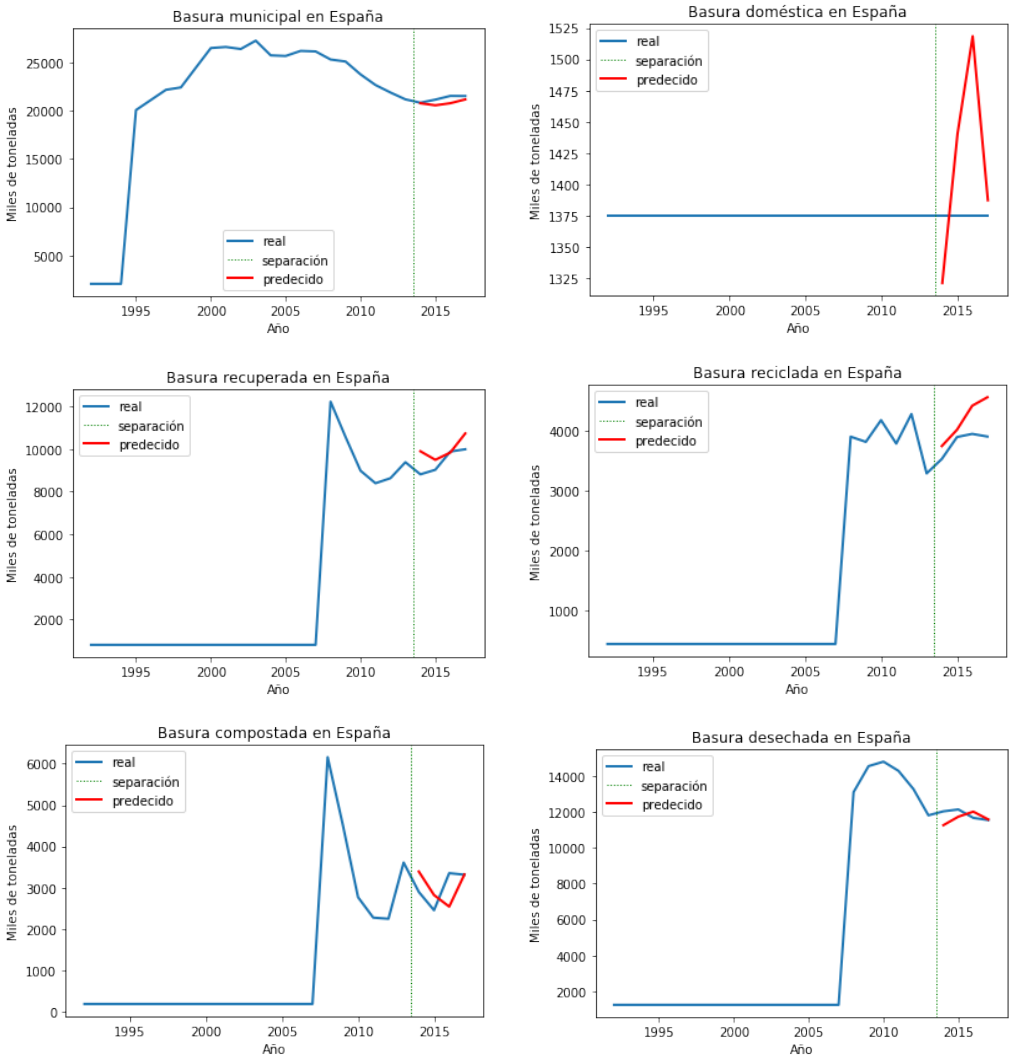

Figura 4.6: Valores de basura reales (azul) y predicho (rojo) para los 6 tipos definidos en España utilizando los GBR definidos en la tabla 4.6.

4.7. Experimento 7: LSTM y DNN

Como últimos experimentos probaremos distintos tipos de redes neuronales. En primer lugar, probaremos con una LSTM. Se realizaron varias pruebas y con la que mejor coeficiente R^2 se obtuvo fue con una LSTM con la siguiente estructura:

1.– LSTM con 100 neuronas, función de activación ReLU
2.– LSTM con 100 neuronas, función de activación ReLU
3.– salida con 7 neuronas, función de activación ReLU

Pese a que las LSTM están en boca de todos para series temporales, en este caso su resultado ha sido realmente malo, obteniendo un coeficiente R^2 negativo, es decir, que actúa peor que un regresor que devuelva la media de cada variable a predecir. En el caso de la red presentada, su coeficiente R^2 es de $-5{,}5218$.

Pensamos que el mal rendimiento de este método viene motivado por la manera en la que queremos enfocar el tratamiento de los datos en este trabajo. Para poder comparar con el resto de algoritmos, a la LSTM sólo se le aporta 3 años de historia para predecir el valor de basuras del año actual. Por otra parte, no introducimos la serie de las basuras ya que no pretendemos predecir para años posteriores al año actual y de esta manera seguimos el mismo esquema que para los experimentos 4.1, 4.2 y 4.3.

Por último, realizaremos una prueba con una red neuronal con la siguiente estructura:

1.– 200 neuronas, función de activación ReLU
2.– 200 neuronas, función de activación ReLU
3.– 200 neuronas, función de activación ReLU
4.– 200 neuronas, función de activación ReLU
5.– 200 neuronas, función de activación ReLU
6.– 200 neuronas, función de activación ReLU
7.– 200 neuronas, función de activación ReLU
8.– Salida con 7 neuronas, función de activación ReLU

En este caso los resultados son mucho mejores, obteniendo un coeficiente R^2 de $0{,}9526$. La estructura de la red se ha elegido entre una serie de experimentos, aportando los mejores resultados. Dado que este algoritmo no obtiene resultados tan relevantes como los primeros 6 experimentos, nos vamos a centrar en ellos.

4.8. Análisis de los datos mediante SVM sin conocimiento del pasado

Ampliando la experimentación de este trabajo, trataremos de evaluar el experimento 4.1 en otras condiciones, variando la cantidad de años de entrenamiento y comprobando el error para cada país. Esperamos ver que aquellos países con errores distribuidos de manera similar posean unas características similares. Hemos escogido el primer experimento porque consideramos que entrenar con información del año previo aporta demasiada información, tal y como se aprecia en la matriz de covarianzas 3.1. Entre los algoritmos entrenados sin datos del pasado, escogemos el SVM porque es el que mejores resultados obtiene.

En primer lugar, vemos cómo afecta el número de años en el conjunto de entrenamiento, lo que se puede ver en la imagen 4.7. Podemos ver que el error decrece según añadimos más muestras en el *dataset* de entrenamiento. El error se estabiliza a partir del corte en 2006, es decir, introduciendo 16 años para entrenar. Pensamos que este error se reduce por la inclusión de datos al *dataset* por parte de países que no habían reportado su información hasta ese año.

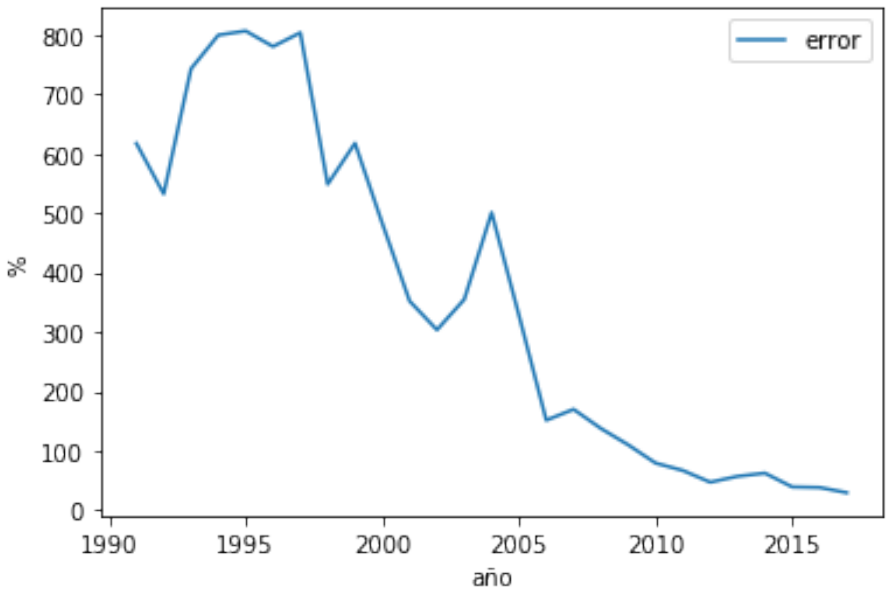

Figura 4.7: Error en porcentaje de un SVR en función del año de corte para separar entrenamiento y prueba.

Para comprobarlo realizaremos cuatro entrenamientos, correspondiendo a cuatro años de corte para los datos (2000, 2005, 2010 y 2015). Con cada modelo generado probaremos a evaluar un país, concretamente Estados Unidos. Utilizaremos sólo las basuras municipales para comprobarlo. En la figura 4.10 podemos apreciar que cuántos más años introducimos, mejor resultado obtenemos, acercando la predicción a los valores reales.

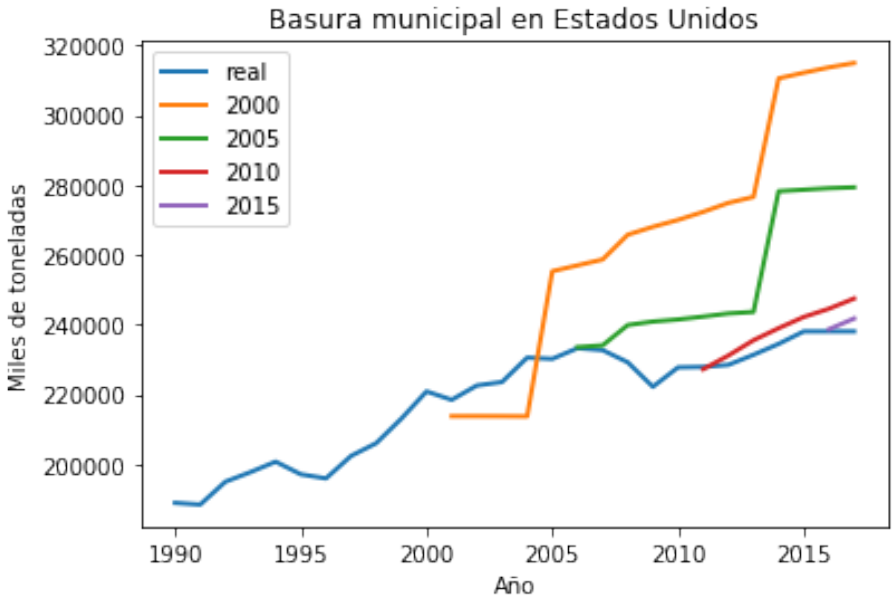

Figura 4.8: Predicciones para las basuras municipales de Estados Unidos en función del año de separación entre datos de entrenamiento y test.

Continuando con este capítulo, utilizaremos los resultados obtenidos con el corte a partir de 2006 para tratar de buscar relaciones entre los países de la OCDE en función al error respecto al modelo. Usaremos la media de errores al predecir los años posteriores al corte de todos los modelos generados para cada corte por encima de 2006. Utilizaremos el porcentaje de error suponiendo que el modelo predice la media de la OCDE para los factores socio-económicos aportados. Vamos a definir los errores como porcentajes; en caso de que dicho porcentaje sea positivo, indicará que el modelo predice un valor mayor que el valor real. En caso contrario, significará que el país produce más basura de la que predice el modelo.

En primer lugar, destacamos que la media del error es $21,66\,\%$, es decir, que de media los países generan menos basura de lo que predicen los modelos. Nos fijaremos en aquellos países que más diferencia tengan respecto al modelo siendo el error positivo. Con un error del $427,20\,\%$ encontramos a Islandia, en segundo lugar Luxemburgo con un $375,26\,\%$ y, en tercer lugar, Canadá con un $143,01\,\%$. Si nos fijamos en las mayores diferencias siendo negativas destaca Estonia con un $-178,58\,\%$, India con un $-164{,}26\,\%$ y Letonia con un $-97{,}27\,\%$. En cuarto lugar encontraríamos a Lituania, situando a las repúblicas bálticas como un conjunto de países que generan más basura que lo que predice el modelo.

India es un país en vías de desarrollo cuya economía provoca un aumento en la generación de basuras, probablemente influido por la instauración de fábricas para empresas internaciones debido a la mano de obra barata. En caso opuesto, vemos que los países que quedan por debajo del modelo son países altamente desarrollados y con un nivel de vida alto. El error provocado en las repúblicas bálticas nos hace pensar que la generación de residuos se puede ver influida por otros factores aparte de los

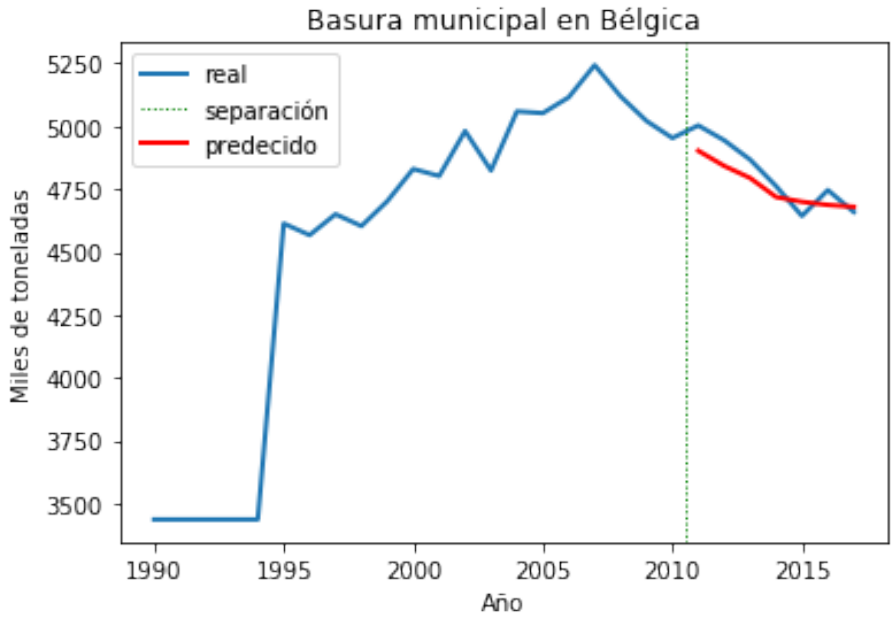

Figura 4.9: Predicción para Bélgica utilizando sólo los datos de dicho país para entrenar el algoritmo del experimento 4.1

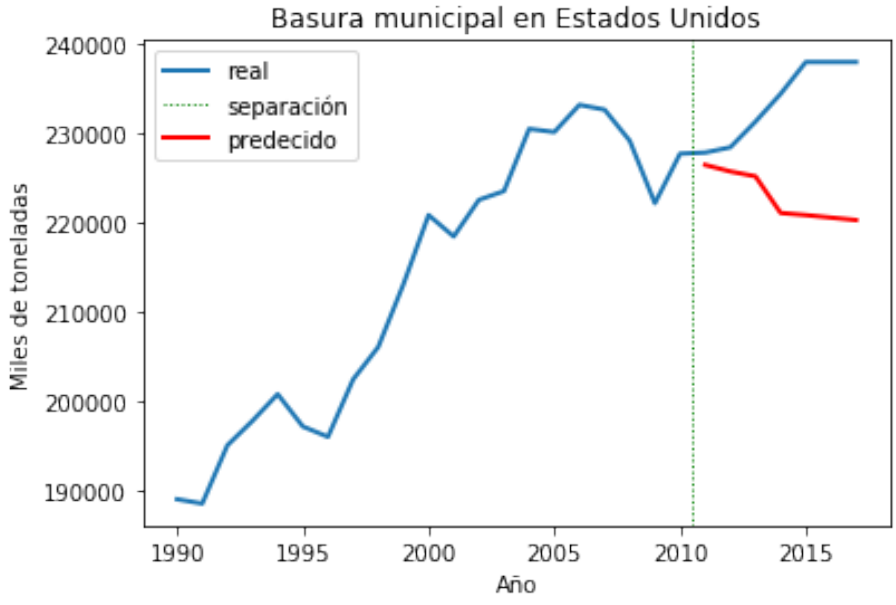

Figura 4.10: Predicción para Estados Unidos utilizando sólo los datos de dicho país para entrenar el algoritmo del experimento 4.1

socio-económicos. Quizá la inclusión de factores culturales o políticos puedan mejorar enormemente las predicciones de esta clase de modelos.

Para finalizar este capítulo, probaremos a reducir la información recibida por el algoritmo utilizado en el experimento 4.1 centrándonos en la basura municipal. Restringiremos los datos a sólo un país y realizaremos el corte en el año 2010. Aquí mostraremos dos países, Bélgica y Estados Unidos, analizaremos el rendimiento prediciendo desde el año 2011 al 2017. Aplicaremos tres métricas: porcentaje de error medio, porcentaje de error en la primera predicción (año 2011) y error en la última predicción (año 2017).

Comenzaremos con Bélgica. En la figura 4.9 se aprecia que la predicción es bastante cercana. El error medio es $1{,}3575\,\%$, el error en la primera predicción es del $2{,}0539\,\%$ y el error en la última es del $0{,}4363\,\%$.

En la figura 4.10 vemos que la predicción para Estados Unidos no es tan cercana como para Bélgica, aumentando el error según se aleja la predicción del corte. El error medio es del $4{,}9089\,\%$, el error en la primera predicción es del $0{,}6129\,\%$ y el error en la predicción más alejada es del $8{,}0586\,\%$.

En este experimento vemos que el error es menor que entrenando con todos los países juntos. Pensamos que esta mejora se debe a que los datos del mismo país mantienen la misma relación entre los factores socio-económicos, eliminando la variabilidad existente entre países. El aumento en el error en Estados Unidos se puede deber a la distancia al punto de corte, en un país de cambio tan rápido como Estados Unidos la relación entre factores socio-económicos y la generación de basuras puede verse modificada progresivamente.

5

Conclusiones

Este TFG tenía como objetivo predecir la generación de residuos sólidos evaluando un conjunto de algoritmos diferentes, así como diferentes técnicas de formar el *dataset*. Hemos visto que introducir información del año anterior mejoraba el resultado de la predicción, consiguiendo valores realmente cercanos a los reales.

En ambos casos (con y sin conocimiento del pasado), el algoritmo de regresión mediante vectores de soporte (SVR) es el que mejores resultados obtiene. El proceso Gaussiano obtiene resultados muy cercanos al SVR. En el caso de las redes neuronales profundas secuenciales, obtienen un resultado muy similar al *boosting* por gradiente. Estos resultados son comparables a los recogidos en [29], donde se muestra el uso de estos mismos algoritmos bajo diversas circunstancias con buenas predicciones. Por último, las redes neuronales LSTM no parecen ofrecer un resultado satisfactorio en este problema.

Prácticamente todos los datos se han obtenido de la página de estadísticas de la OCDE y entre ellos se incluyen variables demográficas, como la población total o algunos grupos de edad, variables geográficas, como la proporción de área construida y variables económicas como la mediana de ingresos. Algunos de estos datos sólo se recogen en periodos de tiempo diferentes al año, en este caso se ha extrapolado la información que faltaba.

Los datos de la OCDE respecto a la generación de basuras son bastante pobres. Algunos países no comienzan a reportar la información solicitada hasta varios años después del comienzo de la recopilación. Esto se puede ver en el caso de España, donde se comienza a reportar 4 de las 6 variables en 2008, una de ellas en 1995 y la última no ha sido reportada en ningún momento. En todos casos, la basura municipal es la que más datos presenta y, por ello, se ha utilizado para tratar de buscar una relación entre países.

Buscando esta relación vemos la similitud entre las repúblicas bálticas y la India, país en vías de desarrollo, así como una relación entre países altamente desarrollados y con gran nivel de vida como Islandia, Canadá y Luxemburgo.

De igual manera, pensamos que la falta de variables socio-culturales o políticas pueden influenciar en gran medida las predicciones de estos modelos por la influencia que tienen en estos factores, donde

el reciclaje depende del presupuesto asignado al mismo y la cultura de consumo y reciclaje.

La implantación de un sistema similar para predecir la generación de residuos de un país para destinar recursos de manera eficiente y, por tanto, evitar el desecho descontrolado al mar o vertederos que habitualmente se encuentran cerca de núcleos urbanos podría marcar la diferencia en calidad de vida de muchos países. En ese caso se podría generalizar, de manera que se usen los datos de provincias o localidades permitiendo una mayor segmentación de la información y, por tanto, una mejor predicción.

Como un posible trabajo posterior se podrían probar otros algoritmos de predicción, así como introducir las componentes políticas y culturales que se comentaban. Por último, se podría probar a ampliar el conjunto de países añadiendo todos los países del mundo y buscando relaciones entre continentes o zonas geográficas. Adicionalmente, otro trabajo futuro podría consistir en agrupar (mediante algún algoritmo de *clustering*) a los países que presenten resultados de predicciones de residuos similares cada año con vistas a definir regiones geográficas comunes y también para entender mejor la influencia de las variables de entrada al sistema en los resultados de predicción.

Todo el código utilizado para este trabajo se encuentra en Github (https://github.com/OscarGB/TFGInf).

Bibliografía

[1] K. Miezah, K. Obiri-Danso, Z. Kádár, B. Fei-Baffoe, and M. Y. Mensah, "Municipal solid waste characterization and quantification as a measure towards effective waste management in Ghana," *Waste Management*, vol. 46, pp. 15–27, 12 2015.

[2] P. Beigl, S. Lebersorger, and S. Salhofer, "Modelling municipal solid waste generation: A review," *Waste Management*, vol. 28, pp. 200–214, 1 2008.

[3] R. Noori, A. Karbassi, and M. Salman Sabahi, "Evaluation of PCA and Gamma test techniques on ANN operation for weekly solid waste prediction," *Journal of Environmental Management*, vol. 91, pp. 767–771, 1 2010.

[4] E. C. Gentil, D. Gallo, and T. H. Christensen, "Environmental evaluation of municipal waste prevention," *Waste Management*, vol. 31, pp. 2371–2379, 12 2011.

[5] A. G. Salman, B. Kanigoro, and Y. Heryadi, "Weather forecasting using deep learning techniques," in *ICACSIS 2015 - 2015 International Conference on Advanced Computer Science and Information Systems, Proceedings*, pp. 281–285, Institute of Electrical and Electronics Engineers Inc., 2 2016.

[6] S. Ramos, S. Gehrig, P. Pinggera, U. Franke, and C. Rother, "Detecting unexpected obstacles for self-driving cars: Fusing deep learning and geometric modeling," in *IEEE Intelligent Vehicles Symposium, Proceedings*, pp. 1025–1032, Institute of Electrical and Electronics Engineers Inc., 7 2017.

[7] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, International Conference on Learning Representations, ICLR, 2015.

[8] M. Abbasi, M. Abduli, B. Omidvar, and A. Baghvand, "Results uncertainty of support vector machine and hybrid of wavelet transform-support vector machine models for solid waste generation forecasting," *Environmental Progress & Sustainable Energy*, vol. 33, pp. 220–228, 4 2014.

[9] D. Antanasijević, V. Pocajt, I. Popović, N. Redžić, and M. Ristić, "The forecasting of municipal waste generation using artificial neural networks and sustainability indicators," *Sustainability Science*, vol. 8, pp. 37–46, 1 2013.

[10] "(PDF) Municipal Solid Waste Data Quality on Artificial Neural Network Performance."

[11] M. Purcell and W. L. Magette, "Prediction of household and commercial BMW generation according to socio-economic and other factors for the Dublin region," *Waste Management*, vol. 29, pp. 1237–1250, 4 2009.

[12] D. Basak, S. Pal, and D. Patranabis, "Support Vector Regression," *Neural Information Processing – Letters and Reviews*, vol. 11, 2 2007.

[13] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics and Data Analysis*, vol. 38, pp. 367–378, 2 2002.

[14] A. J. Smola· and P. Bartlett, “Sparse Greedy Gaussian Process Regression,” tech. rep.

[15] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” 5 2015.

[16] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, “LSTM: A Search Space Odyssey,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, pp. 2222–2232, 10 2017.

[17] F. Pedregosa, G. Varoquaux, and A. Gramfort, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[18] A. Gulli and S. Pal, “Deep learning with Keras,” 2017.

[19] Martin Abadi, Ashish˜Agarwal, Paul˜Barham, Eugene˜Brevdo, Zhifeng˜Chen, Craig˜Citro, Greg˜S.˜Corrado, Andy˜Davis, Jeffrey˜Dean, Matthieu˜Devin, Sanjay˜Ghemawat, Ian˜Goodfellow, Andrew˜Harp, Geoffrey˜Irving, Michael˜Isard, Y. Jia, Rafal˜Jozefowicz, Lukasz˜Kaiser, Manjunath˜Kudlur, Josh˜Levenberg, Dan˜Mané, Rajat˜Monga, Sherry˜Moore, Derek˜Murray, Chris˜Olah, Mike˜Schuster, Jonathon˜Shlens, Benoit˜Steiner, Ilya˜Sutskever, Kunal˜Talwar, Paul˜Tucker, Vincent˜Vanhoucke, Vijay˜Vasudevan, Fernanda˜Viégas, Oriol˜Vinyals, Pete˜Warden, Martin˜Wattenberg, Martin˜Wicke, Yuan˜Yu, and Xiaoqiang˜Zheng, “TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems,” 2015.

[20] E.-A. N. Antonio Gallardo, P. D. Professor, M. Carlos, F. J. Colomer, A. M. Gómez-Parra, P. D. Researcher, and N. Edo-Alcon, “INFLUENCE OF THE INCOME FACTOR IN THE MUNICIPAL SOLID WASTE SELECTIVE COLLECTION,” tech. rep.

[21] N. J. Bandara, J. P. A. Hettiaratchi, S. C. Wirasinghe, and S. Pilapiiya, “Relation of waste generation and composition to socio-economic factors: A case study,” *Environmental Monitoring and Assessment*, vol. 135, pp. 31–39, 12 2007.

[22] P. T. T. Trang, H. Q. Dong, D. Q. Toan, N. T. X. Hanh, and N. T. Thu, “The Effects of Socio-economic Factors on Household Solid Waste Generation and Composition: A Case Study in Thu Dau Mot, Vietnam,” in *Energy Procedia*, vol. 107, pp. 253–258, Elsevier Ltd, 2 2017.

[23] M. Awad, R. Khanna, M. Awad, and R. Khanna, “Support Vector Regression,” in *Efficient Learning Machines*, pp. 67–80, Apress, 2015.

[24] J. Elith, J. R. Leathwick, and T. Hastie, “A working guide to boosted regression trees,” *Journal of Animal Ecology*, vol. 77, pp. 802–813, 7 2008.

[25] “The Rise and Fall of the G.D.P. - The New York Times.”

[26] P. S. Morrison and B. Beer, “Consumption and Environmental Awareness: Demographics of the European Experience,” pp. 81–102, Springer, Singapore, 2017.

[27] W. McKinney, “pandas: a foundational Python library for data analysis and statistics,” *Python for High Performance and Scientific Computing*, vol. 14, 2011.

[28] S. K. Patro and K. K. sahu, “Normalization: A Preprocessing Stage,” *IARJSET*, pp. 20–22, 3 2015.

[29] K. Kolekar, T. Hazra, and S. Chakrabarty, “A Review on Prediction of Municipal Solid Waste Generation Models,” *Procedia Environmental Sciences*, vol. 35, pp. 238–244, 1 2016.


UAM
UNIVERSIDAD AUTONOMA
DE MADRID